Claims

We claims:

1. A method for managing traffic over a channel of a network connecting a sender end system and a receiver end system, the traffic including a plurality of packets, comprising:

modeling the channel as a queue having an associated queue occupancy;

maintaining, in the sender end system, a series of times when packets are sent and a series of times when feedback messages are received by the sender system in response to receiving the feedback messages in the sender end system indicating the packets were received by the receiver end system;

updating a time series of samples for a service time experienced by each packet sent based on the series of times when the packets are sent and the feedback messages are received;

predicting a most recent queue occupancy based on the time series of samples; and

sending the next packet according to the predicted queue occupancy.

2. The method of claim 1 further comprising

sending the next packet immediately if the queue occupancy is less than one; otherwise

sending the next packet when the feedback message is received for a current packet if the queue occupancy is one; and otherwise

delaying sending the next packet until the queue occupancy is one.

3. The method of claim 1 wherein the predicting uses a multi-timescale linear prediction method.

4. The method of claim 1 wherein each sample t_s equals a departure time of a packet n - maximum(departure time of a packet n-1, arrival time of the packet n).

5. The method of claim 3 further comprising:
subtracting a mean μ for the time series from each pair of samples to produce a zero-mean time series for the predicting.

6. The method of claim 1 further comprising:

counting lost packets;

inferring and updating the available queue occupancy

considering the lost packets when predicting the queue occupancy; and

using the available queue occupancy to determine a speed of congestion control.

7. The method of claim 1 wherein the available queue occupancy is used to predict packet loss and to inform an encoder.

8. The method of claim 1 wherein the sender end system is connected to the receiver end system via a relay, and the channel includes a link from the sender end system to the relay and a link from the relay to the receiver end system.

9. The method of claim 8 further comprising:

predicting a first queue occupancy for a first link at the sender;

predicting a relay buffer fullness at the sender; and

predicting a second queue occupancy for a second link at the relay.

10. The method of claim 1 wherein each feedback message includes application feedback data and transport feedback data.

11. The method of claim 1 further comprising:

reducing a number of feedback messages sent by the receiver end system when the predicted queue occupancy is within a predetermined error measure.

12. The method of claim 8 wherein the relay includes independently operating traffic management and content adaptation modules.

13. The method of claim 8 wherein the sender and the relay form a first control loop, and the relay and the receiver form a second control loop.

14. The method of claim 12 wherein the content adaptation module withdraws over-allocation at the relay when the sender over-allocates bandwidth.

15. The method of claim 14 wherein the sender updates a total bits allocated based on an over-allocation withdrawal at the relay.

16. A system for managing traffic over a channel of a network connecting a sender end system and a receiver end system, the traffic including a plurality of packets, comprising:

means for modeling the channel as a queue having an associated queue occupancy;

means for maintaining, in the sender end system, a series of times when packets are sent and a series of times when feedback messaged are received by the sender system in response to receiving the feedback messages in the sender end system indicating the packets were received by the receiver end system;

means for updating a time series of samples for a service time experienced by each packet sent based on the series of times when the packets are sent and the feedback messages are received;

means for predicting a most recent queue occupancy based on the time series of samples; and

means for sending the next packet according to the predicted queue occupancy.